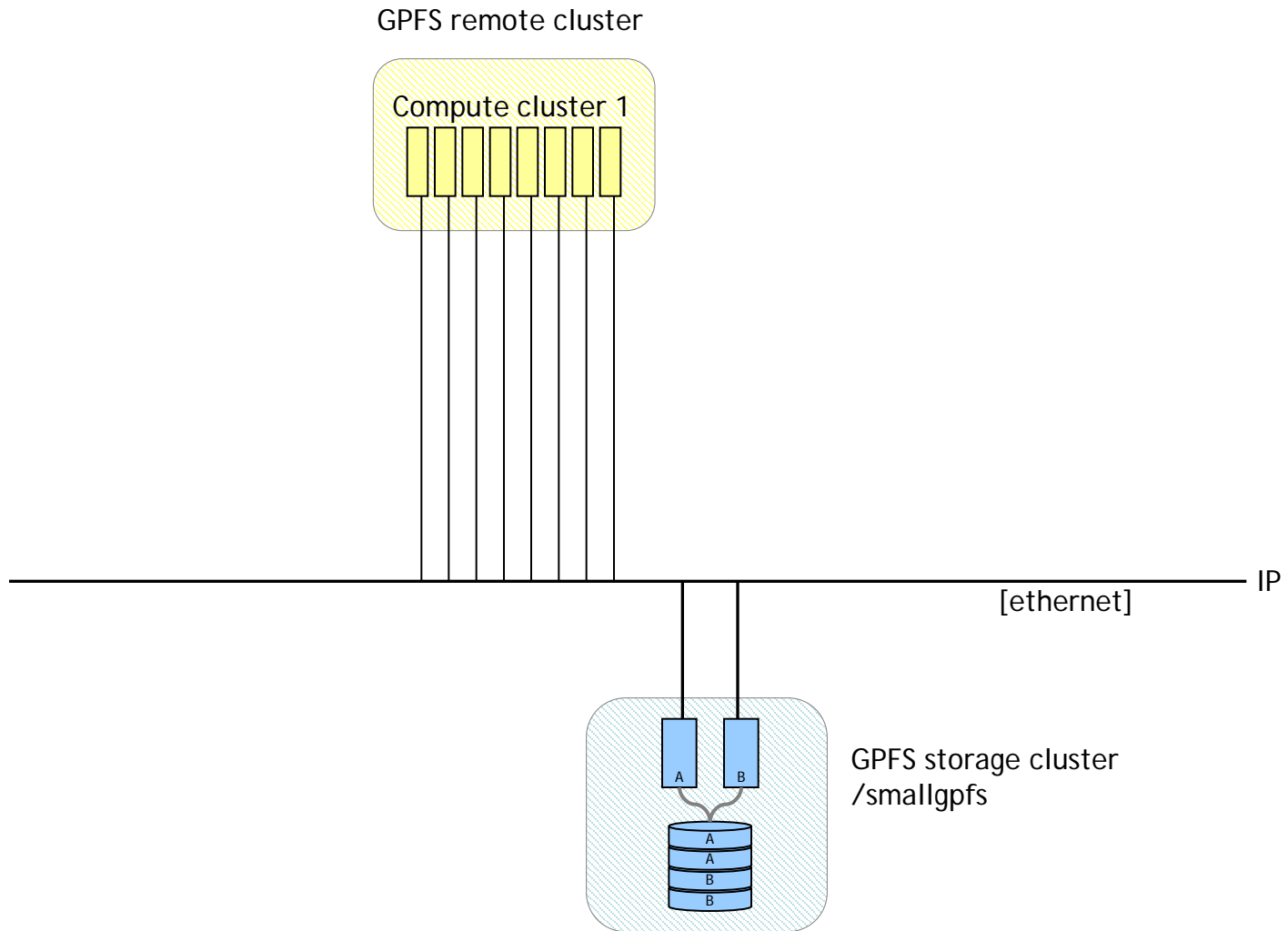


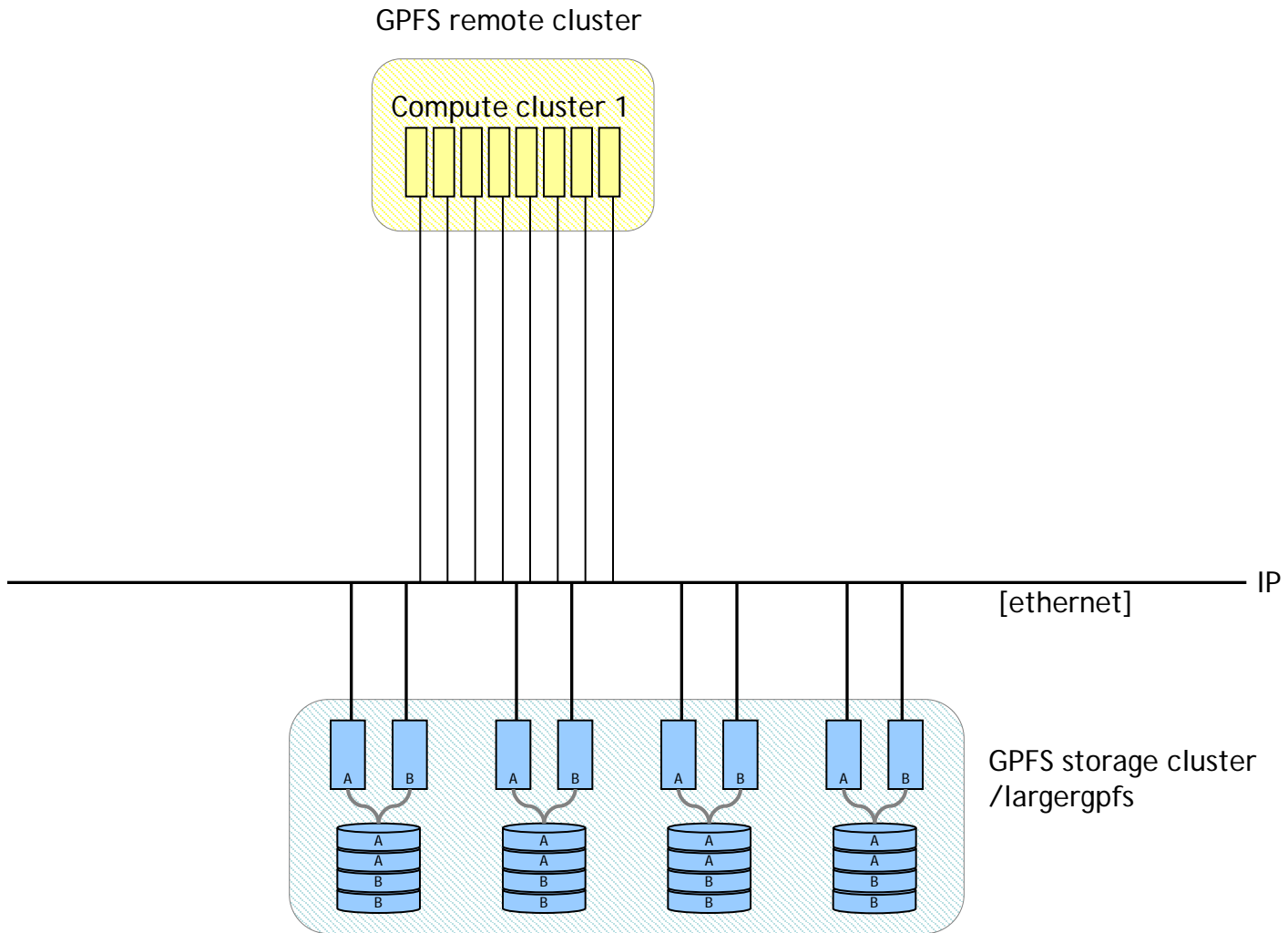
A different approach to Infiniband as the global storage network

Kim Petersen
HPC Solutions Architect

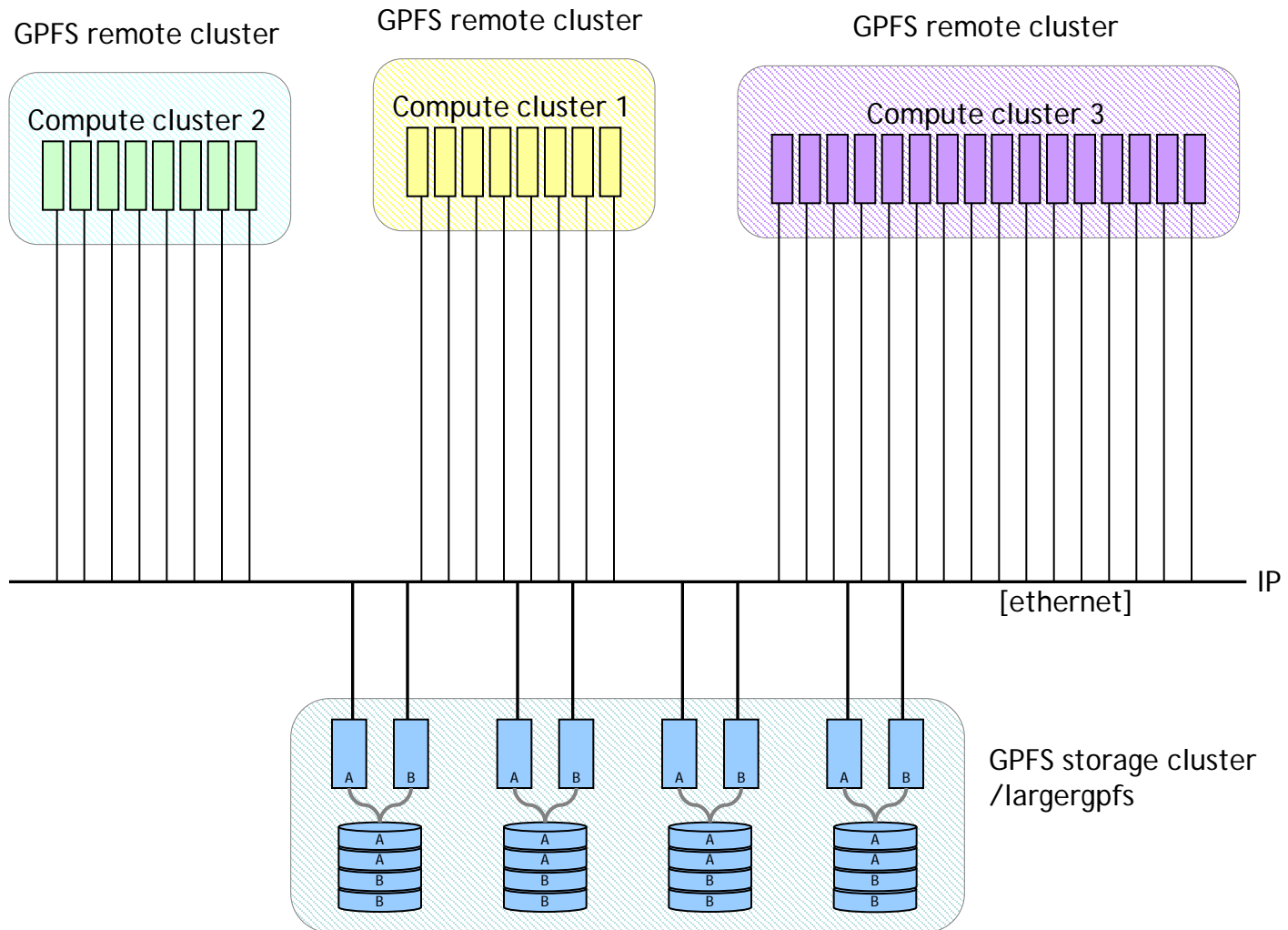
GPFS made simple



GPFS made larger

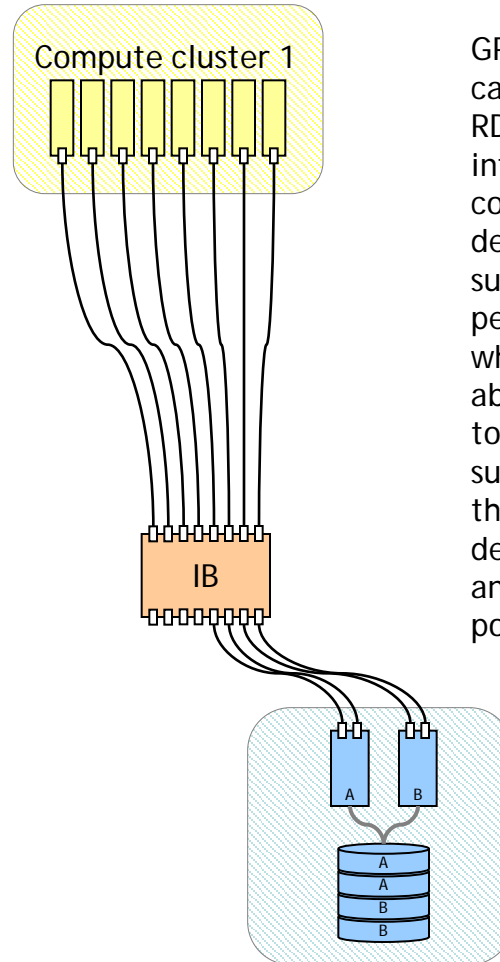


GPFS made more complex



Using a high bandwidth network

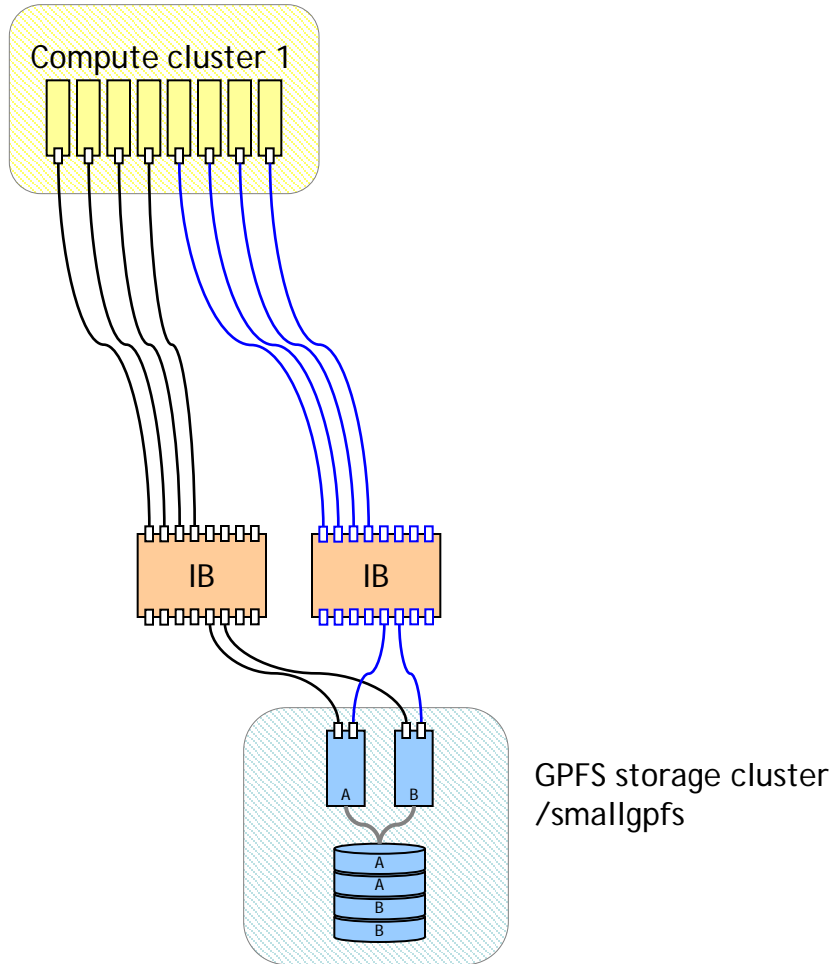
GPFS remote cluster



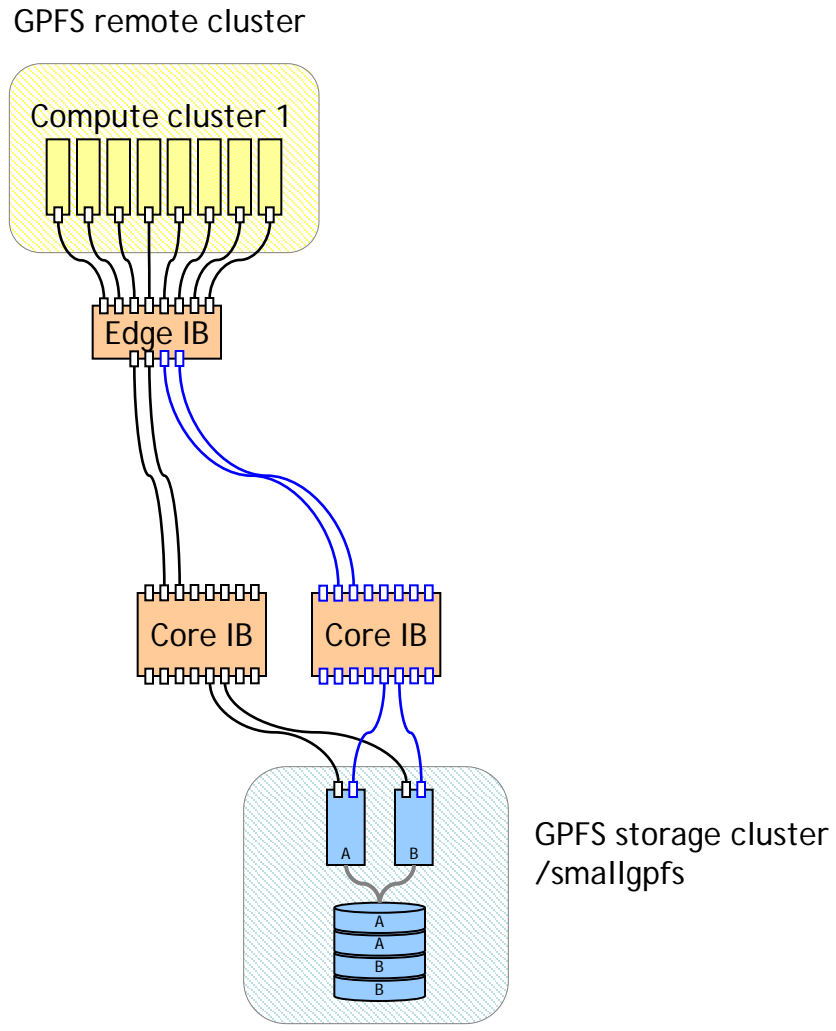
GPFS uses the NSD protocol over any TCP/IP capable network fabric or on Linux it can use a RDMA InfiniBand protocol to transfer control information and data to NSD clients. These communication interfaces need not be dedicated to GPFS; but should provide sufficient bandwidth to meet your GPFS performance expectations and for applications which share the bandwidth. GPFS has the ability to define a preferred network subnet topology, for example designate separate IP subnets for intra-cluster communication and the public network. This provides for a clearly defined separation of communication traffic and allows you to increase the throughput and possibly the number of nodes in a GPFS cluster.

Add some redundancy

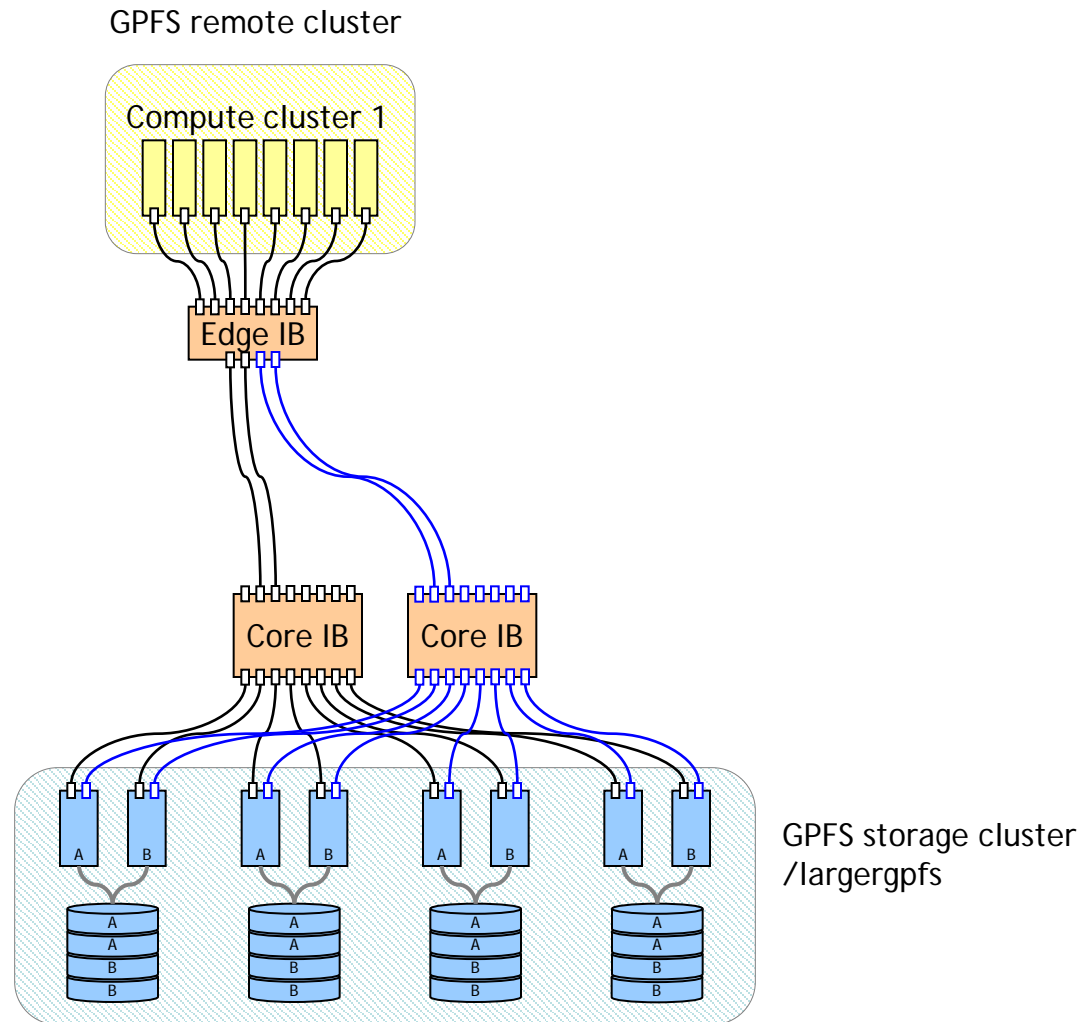
GPFS remote cluster



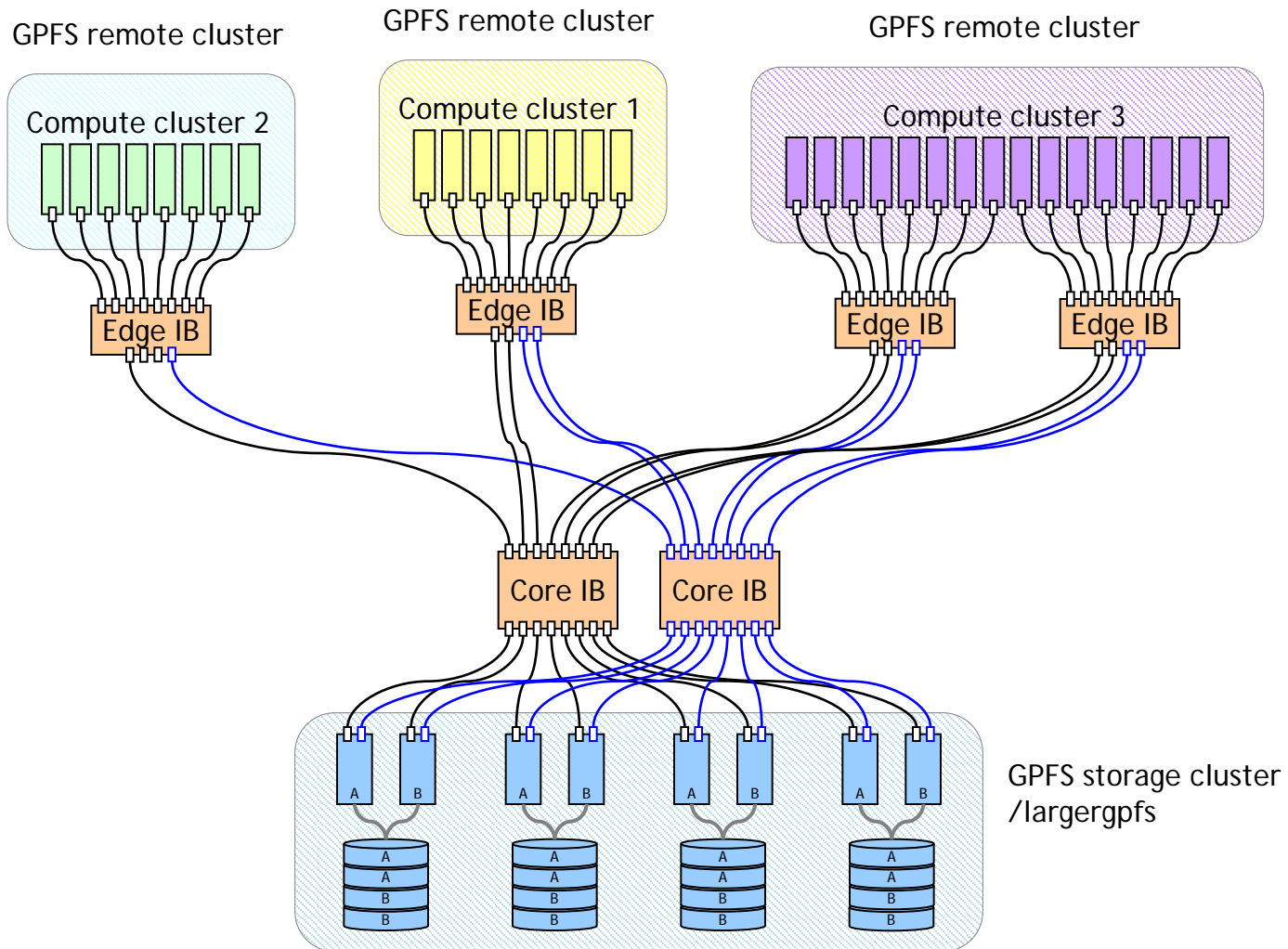
Fault tolerance at the core level



Grow the storage cluster



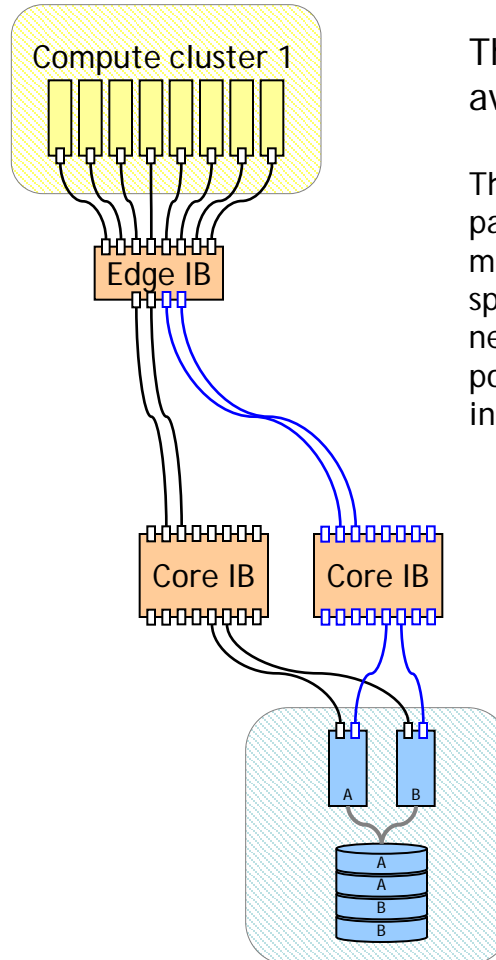
Add remote clusters



So much for the theory ...

(or “anything goes in PowerPoint but what about real life?”)

GPFS remote cluster



The GPFS storage cluster **MUST** be available and accessible at all times.

This requires IP link aggregation (bonding) as part of IP over IB (IPoIB). A GPFS (NSD) server must aggregate (bond) the two IB ports in IP space to provide a single IP interface for the IP network! Needed elsewhere (where two links possible per server) to provide a single IP interface.

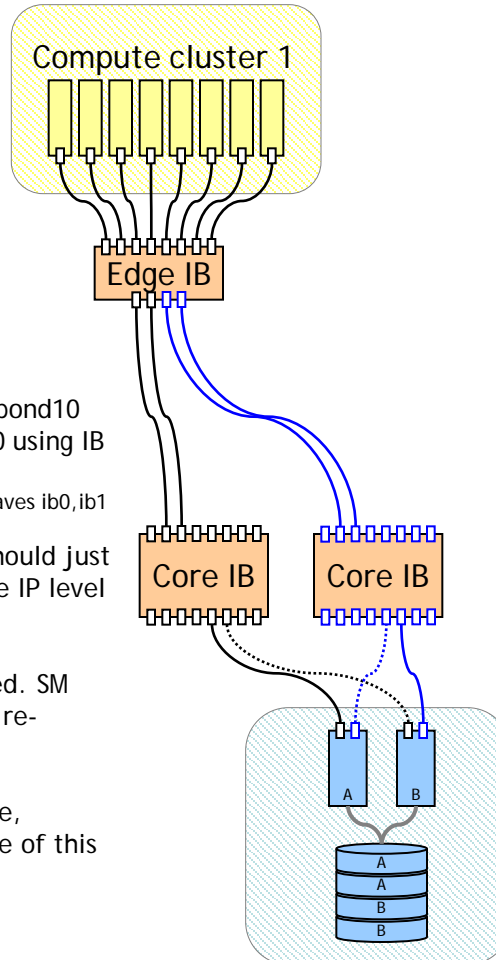
HA bonding

www.openfabrics.org

- High Availability for IPoIB is achieved through the Linux Bonding driver
- The Linux bonding code was changed by Voltaire to support IPoIB
 - basically the change was to allow bonding to use the HW address of the active slave, as with IPoIB one can't assign the HW address (GID, QPN)
- Bonding provides HA at the network stack Link (L2) level
- Basically, layer separation means that TCP sessions should not break.
- HW failure would cause the IB RC session of a native IB ULPs (SDP, RDS, iSER, Lustre, rNFS) to break
 - bonding allows for a new session to be established immediately (as ipoib is the IB stack [rdma_cm] ARP provider)
 - depending on the ULP, this session breakage may not be even seen by the user!
 - Latest improvements (OFED 1.4) in bonding and RDMA CM improves also IB ULPs HA.
- Bonding HA mode
 - called Active-Backup (not for bandwidth aggregation)
 - has one active slave, applies link detection mechanisms to trigger fail-over
 - one HW (L2) address is used for the bond
 - typically the one of the first slave, which is then assigned to the other slaves
- Link detection mechanisms
 - local: uses the carrier bit of the slaves
 - path validation: implemented through an ARP target to which probes are sent
- Fail-over
 - bonding sends a Broadcast Gratuitous ARP (originally to update the Ethernet switches tables)
 - bonding does a "replay" of multicast join
 - latest improvements: Sends net event to RDMA CM → RDMA CM notifies IB ULP / application

The real life ...

GPFS remote cluster



As the bonding HA mode provides an active and a passive (stand-by) connection the configuration of active and passive links becomes important for load-balancing across the redundant core switches.

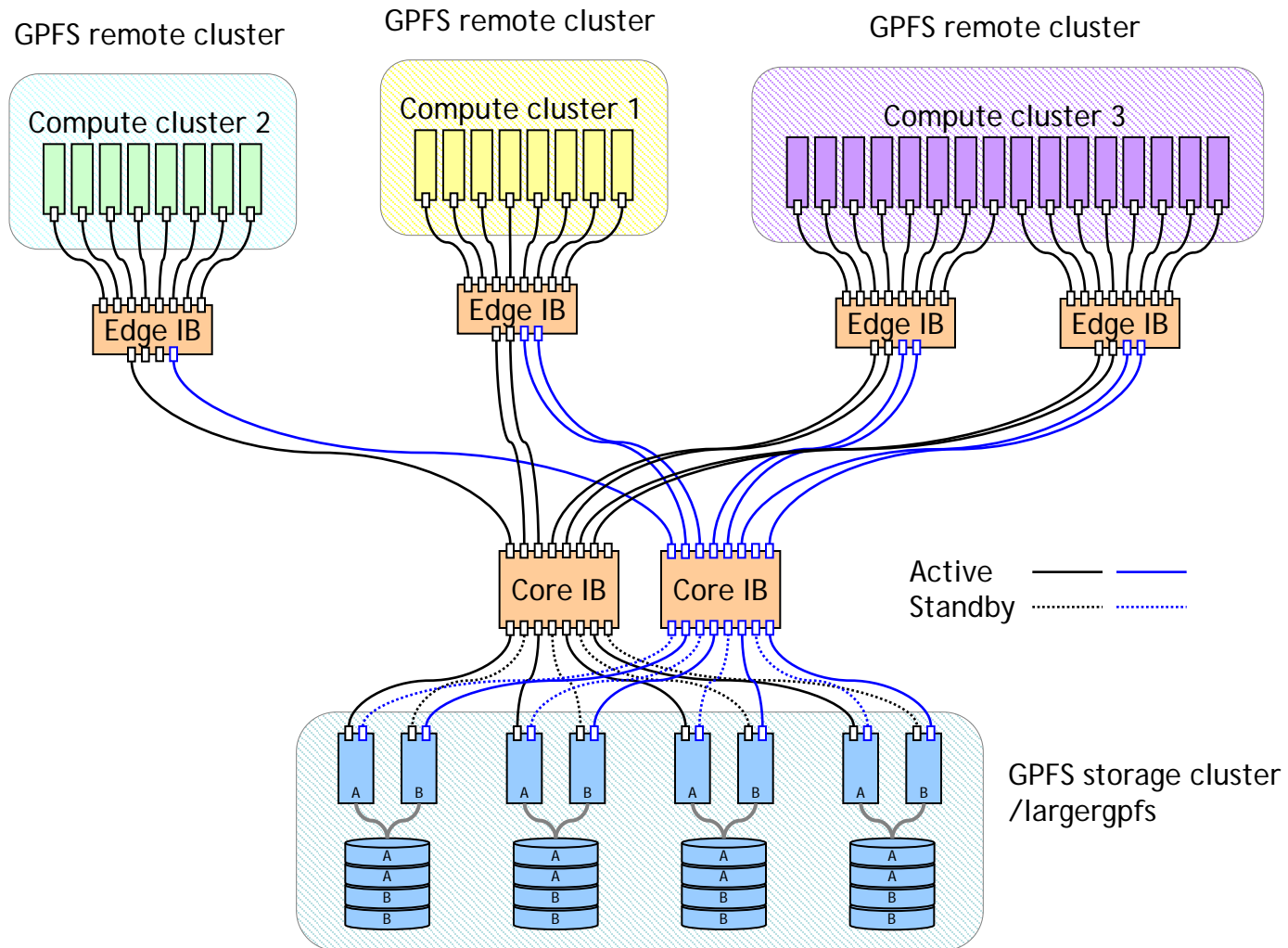
IB-bonding (OFED 1.4); e.g., configure interface bond10 with address 192.168.10.10, netmask 255.255.0.0 using IB interfaces (slaves) ib0, ib1:
ib-bond -bond-name bond10 -bond-ip 192.168.10.10/16 -slaves ib0,ib1

If one core IB switch goes out the GPFS servers should just activate the standby IB port; should not hit at the IP level hence GPFS will not react to this.

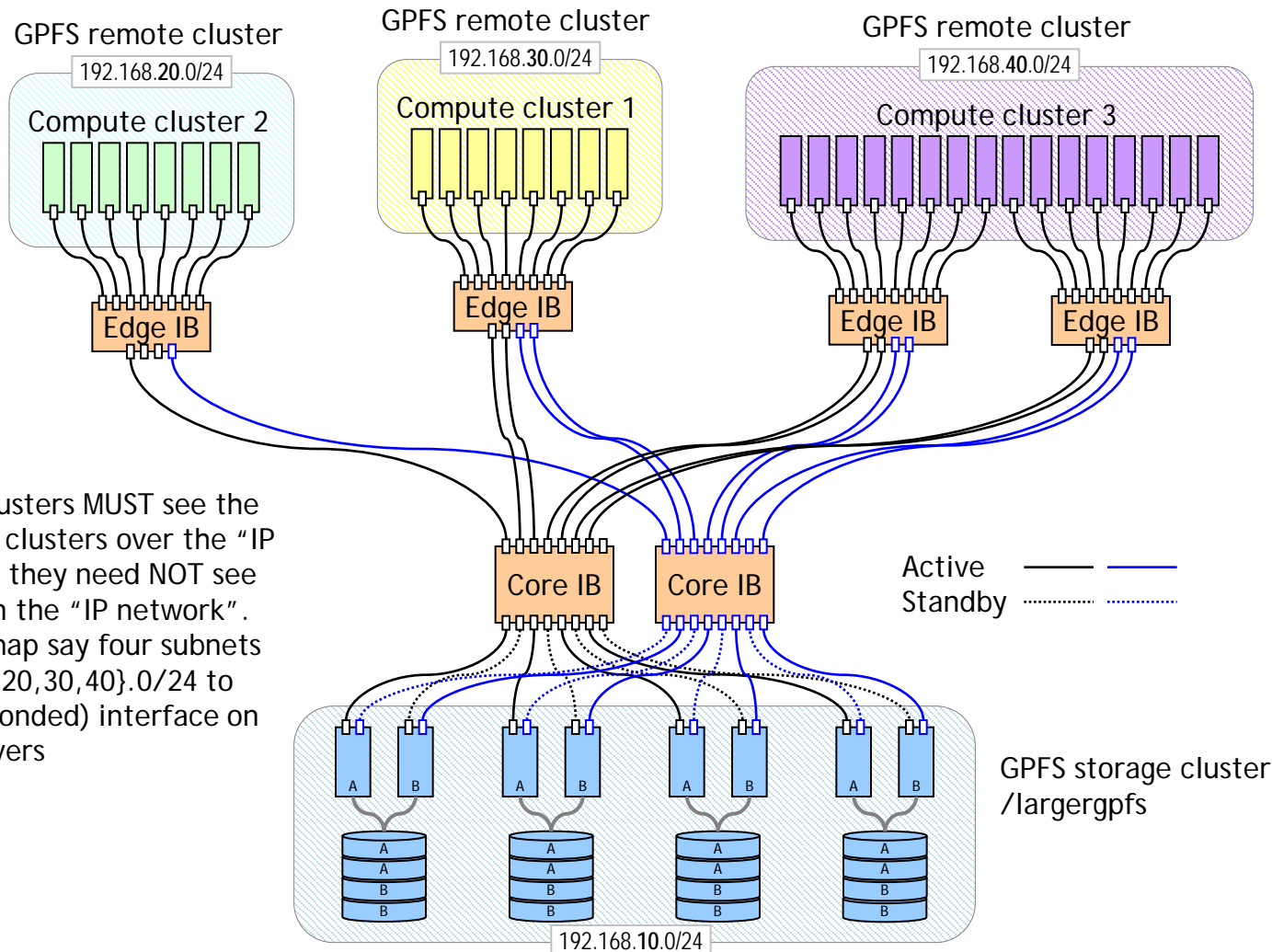
All other IPoIB "paths" should work un-interrupted. SM (subnet manager) will re-route port-to-port links re-establishing IP (TCP sessions should not break).

There are finer details to be considered (MTU size, connection mode, etc) but that is not the purpose of this presentation.

Grow and add remote clusters

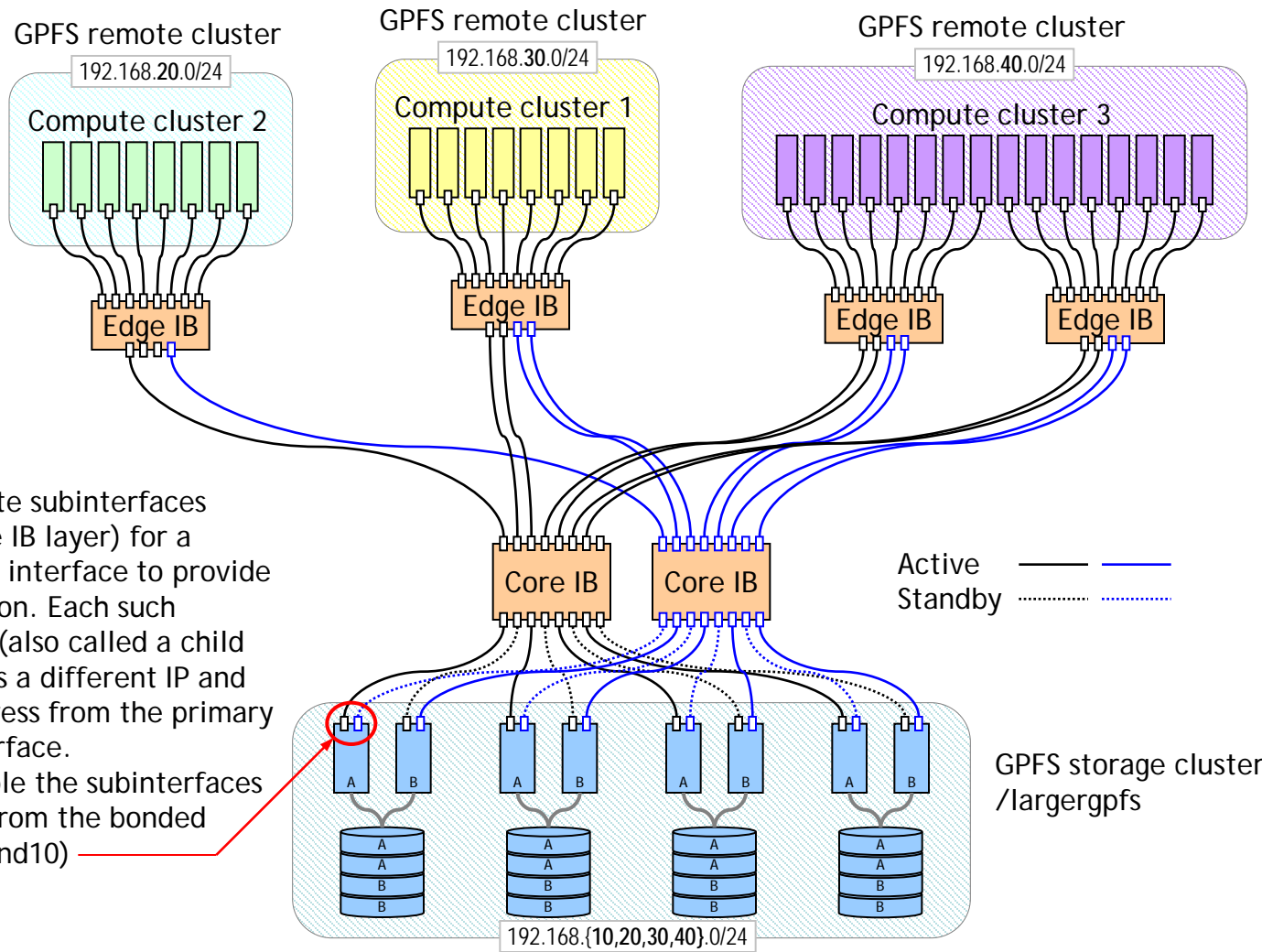


Multiple IP networks; how to ...



All remote clusters MUST see the GPFS storage clusters over the "IP network" but they need NOT see each other on the "IP network". We need to map say four subnets (192.168.{10,20,30,40}.0/24 to the virtual (bonded) interface on the GPFS servers

The sub-interface

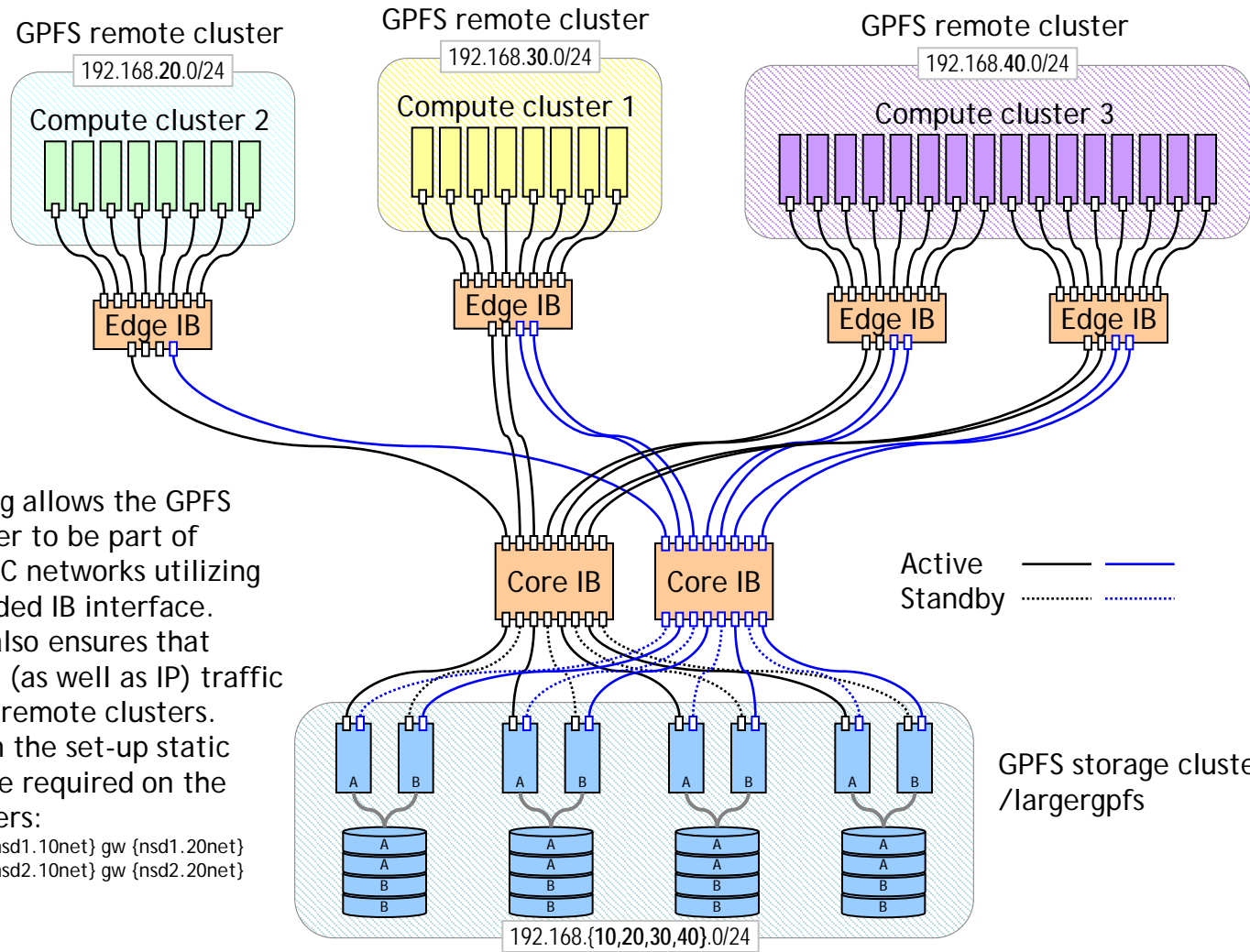


Need to create subinterfaces (partition the IB layer) for a primary IPoIB interface to provide traffic isolation. Each such subinterface (also called a child interface) has a different IP and network address from the primary (parent) interface. In this example the subinterfaces are created from the bonded interface (bond10)

Active ————
Standby ······

GPFS storage cluster /largergpfs

IB partitioning



IB partitioning allows the GPFS storage cluster to be part of several class C networks utilizing a single, bonded IB interface. Partitioning also ensures that there is no IB (as well as IP) traffic between the remote clusters. Depending on the set-up static routes may be required on the remote clusters:

```
# route add -host {nsd1.10net} gw {nsd1.20net}
# route add -host {nsd2.10net} gw {nsd2.20net}
```